16th ICCRTS

"Collective C2 in Multinational Civil-Military Operations"

Semantic Analysis of Military Relevant Texts for Intelligence Purposes

Topic 3: Information and Knowledge Exploration

Topic 4: Information and Knowledge Exploitation

Topic 8: Architectures, Technologies, and Tools

Sandra Noubours (point of contact)
Dr. Matthias Hecking

Fraunhofer Institute for Communication,
Information Processing and Ergonomics FKIE
Neuenahrer Straße 20
53343 Wachtberg
Germany

Phone:+49 228 9435 752
Fax: +49 228 9435 685
Email: sandra.noubours@fkie.fraunhofer.de

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Semantic Analysis of Military Relevant Texts for Intelligence Purposes	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Fraunhofer Institute for Communication,Information Processing and Ergonomics FKIE,Neuenahrer Stra? 20,53343 Wachtberg, Germany,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the 16th International Command and Control Research and Technology Symposium (ICCRTS 2011), Qu?c City, Qu?c, Canada, June 21-23, 2011. U.S. Government or Federal Rights License.

14. ABSTRACT
The current deployments of the German Federal Armed Forces cause the necessity to analyze large quantities of intelligence reports and other documents written in different languages. To efficiently handle these tasks natural language processing techniques (NLP) can be applied. The ZENON project makes use of an information extraction approach for the (partial) content analysis of English HUMINT reports. It has further been extended to do multilingual information extraction, i.e., processing Dari and Tajik texts. The focus of this paper is on the improvement of ZENON?s English semantic analysis. Intelligence reports are characterized by a large topical and linguistic variety. In order to extend the system?s coverage when performing content analysis we realized a semantic role labeling approach. In this paper, after a short introduction, the ZENON system and its information extraction functionalities are explained. Then our semantic role labeling approach and the architecture of the implemented application are described in detail.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**42**	

Semantic Analysis of Military Relevant Texts for Intelligence Purposes

Sandra Noubours, Dr. Matthias Hecking

Fraunhofer FKIE
Neuenahrer Straße 20
53343 Wachtberg, Germany

Phone: +49 228 9435 752
Fax: +49 228 9435 685
Email: sandra.noubours@fkie.fraunhofer.de

Abstract

The current deployments of the German Federal Armed Forces cause the necessity to analyze large quantities of intelligence reports and other documents written in different languages. To efficiently handle these tasks natural language processing techniques (NLP) can be applied. The ZENON project makes use of an information extraction approach for the (partial) content analysis of English HUMINT reports. It has further been extended to do multilingual information extraction, i.e., processing Dari and Tajik texts. The focus of this paper is on the improvement of ZENON's English semantic analysis. Intelligence reports are characterized by a large topical and linguistic variety. In order to extend the system's coverage when performing content analysis we realized a semantic role labeling approach. In this paper, after a short introduction, the ZENON system and its information extraction functionalities are explained. Then our semantic role labeling approach and the architecture of the implemented application are described in detail.

1 Introduction

Military intelligence opens up various applications for *natural language processing* (NLP) [1; 2]. Currently, the deployments of the German Federal Armed Forces (Bundeswehr) cause the necessity to analyze large quantities of intelligence reports and other documents written in different languages. NLP techniques can be adopted to efficiently handle these tasks. We set up the research project ZENON in which a NLP approach is used for the (partial) content analysis of free-form texts.

The ZENON project [3; 4] realizes a prototypical information extraction (IE) system to semantically analyze documents. Information about actions and entities that are described in a text are identified. This information, completed with location and time data, is combined into a graphically navigatable Entity-Action-Network (e.g.; with a person in the center of the network). The overall objective of the project is to demonstrate that it is possible to use state-of-the-art natural language processing techniques to extract and combine military relevant knowledge from free-form texts. An expected advantage of systems like ZENON is the increased productivity of intelligence analysts. They might analyze and combine information from a larger volume of intelligence reports and from more open sources as well as in foreign languages. With this assistance the intelligence analysis can be handled in a more efficient way than without such automatic support.

In this paper we describe a *semantic role labeling* (SRL) approach [5] to extend ZENONs content analysis of English text. In the context of a specific action, each entity involved holds a certain *semantic role* that denotes the function of that entity in the course of that action. For example in sentence (1.1) *arrested* is the verb indicating the action. In the course of that event, the entity *the policemen* takes the role of an AGENT (an intentionally acting instance) while *the suspect* is the PATIENT (a person affected by an action).

Sentence (1.1): [*The policemen*]AGENT ***arrested*** [*the suspect*]PATIENT.

Semantic role labeling is the process of automatically indentifying semantic roles in a text [6]. For each action encoded in a text the participating entities are identified and labeled with semantic roles. SRL is an important ongoing NLP research area. Different approaches exist, many of them applying machine learning, where the system is trained on an annotated corpus. The training corpus should be domain specific. As there is no such corpus existent for the military domain we implemented a non-statistical approach that makes use of a lexical resource. SRL has a number of possible applications, for example machine translation and information extraction [7]. In the course of IE, semantic roles constitute further knowledge about actions and entities. We implemented the SRL application to extent the semantic component of the ZENON system. This is expected to improve the all-over performance of the ZENON. The paper is structured as follows: The ZENON System is described with a focus on its semantic processing. Then our semantic role labeling approach is introduced and the implemented SRL system is explained in detail.

2 The ZENON System

The research system *ZENON* [3; 4] realizes an information extraction approach for the (partial) content analysis of intelligence reports. It is able to process English documents and has further been extended to do *multilingual information extraction*, i.e., functionalities were build to process simple *Dari* texts [8; 9] and a module for processing *Tajik* input [10; 3] has been implemented. Starting with English HUMINT reports from the KFOR deployment of the German Federal Armed Forces [11; 12] we developed the first version of ZENON [13; 14]. Such intelligence reports are characterized by large topical and linguistic variety. Apart from descriptions of conflicts between ethnic groups, tensions between political parties, information about infrastructure problems, etc. there are also reports, which concern events and individuals or other entities. For example statements of the form *A meets B*, *A marries C*, *A shoots B*, etc. contain information about activities/events and the entities involved.

Performing a (partial) content analysis of unrestricted text is the purpose of an *information extraction* (IE) [15; 6] system. Relevant information about a specific entity and/or action in natural language texts is identified, collected, and formalized. To implement an IE system, language-specific resources (lexicons, grammars etc.) and appropriate software (parser, tagger, etc) are necessary. In order to improve the system's performance domain knowledge can be integrated in form of lexical resources. For example, for disambiguation purposes, the word *Leopard* in the lexicon can have the categorical information *tank*. The association between words and semantic information is domain-specific and has to be changed depending on a systems application.

ZENON's IE functionality follows a rule-based approach [13], i.e., shallow rule-based processing is applied (e.g., implementing transducers). The text is analyzed with respect to what is of interest for the application. The main advantage of this approach is robustness when confronted with ungrammatical sentences. The disadvantage is that relevant information may possibly be missed.

For the construction of the ZENON research system *GATE*[1] (General Architecture for Text Engineering) [16; 17] is used. GATE is an infrastructure for developing and deploying NLP software. It offers a lot of tools, which were applied and/or extended to implement the natural language processing parts of ZENON (e.g., morphological analyzer, part-of-speech (POS) tagger, pre-defined transducer to recognize English verbal phrases, chunk-parsing).

[1] Online available at http://gate.ac.uk/.

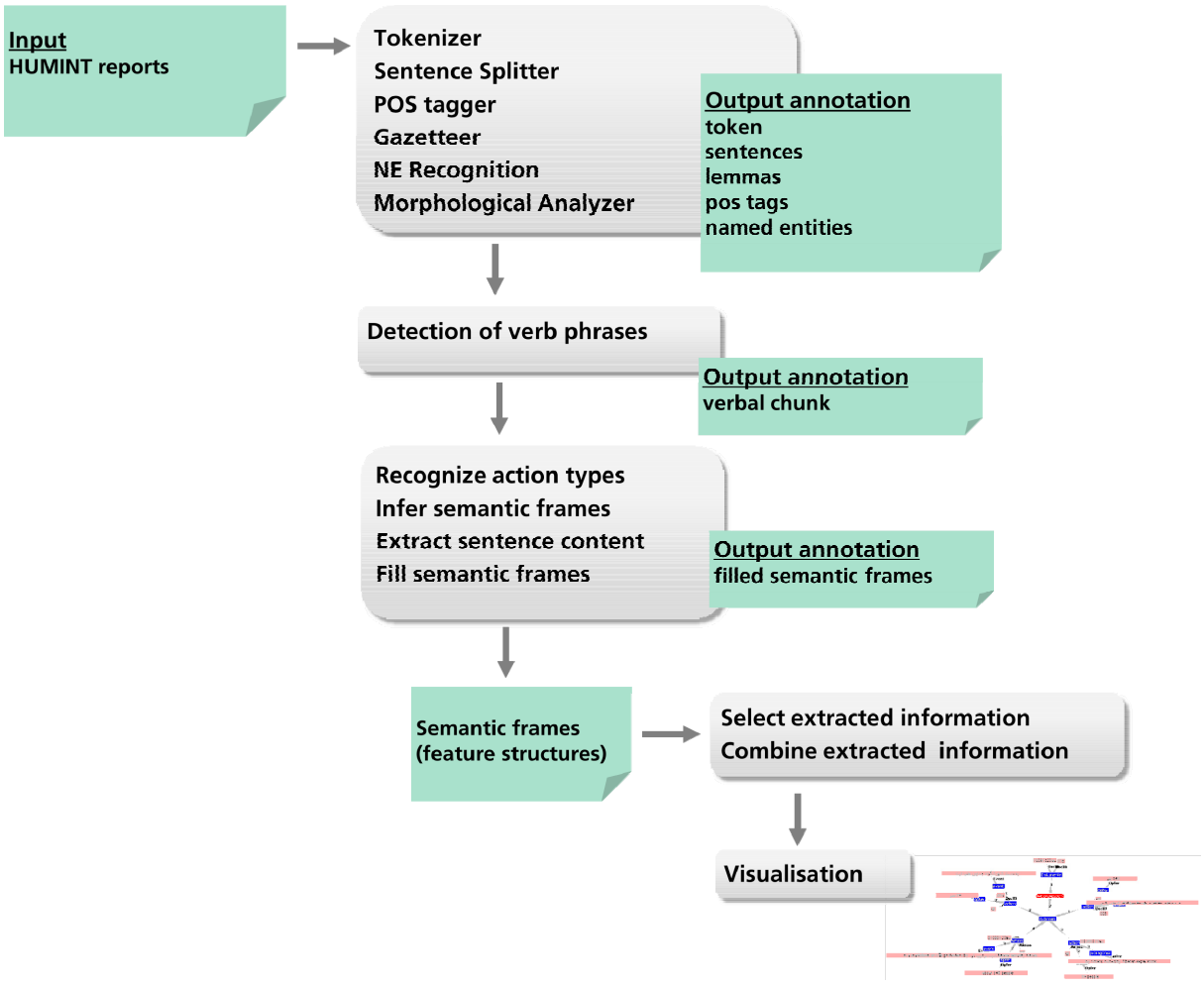


Figure 1: The ZENON processing chain

The system's processing chain is visualized in Figure 1. Natural language text (i.e., HUMINT reports) is fed into the system. The input is analyzed by different processing resources. Each component outputs so called *annotations*, i.e., the information that represents the result of an analysis is noted down together with a specification for what part of text it applies to. In this way the output annotation of a processing resource can be used as input by the next one.

ZENONs processing chain starts with tokenizing the input text, i.e., words, numbers, etc. are detected and sentence boundaries are recognized. For each token the part-of-speech (pos) is determined, that is whether a word is a noun (N), a verb (V), etc. During a morphological analysis each token is annotated with its lemma form (dictionary form).

An important processing step is the recognition of domain- and application-specific named entities (i.e., complex names of e.g., political organizations, person names, etc.). First, simple names of cities, regions, military organizations, etc. are identified (by the Gazetteer). On that basis named entities are determined. In the ZENON prototype transducers for the recognition of following named entities were developed: *City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time* and *Title*.

In the course of ZENON's semantic processing, verb phrases, action types and the sentence content are analyzed. The ZENON prototype uses various transducers for semantic analysis. First, finite and non-finite verbal phrases, modal verb phrases, participles and special composed verb expressions are identified. Based on the detected verb groups, different action types can be determined (e.g., from the infinitive of *murder*, *kill*, *decapitate*, etc. the action class KILL). From the extracted action type an associated *semantic frame* is inferred.

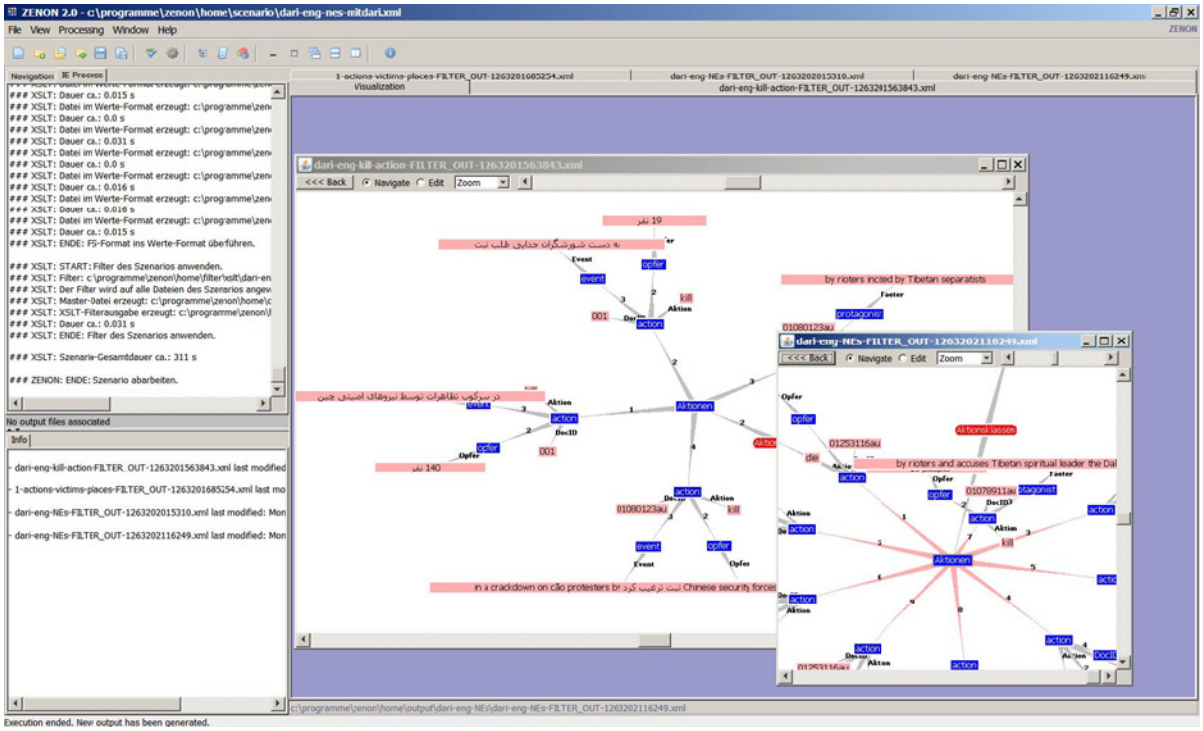


Figure 2: Graphically navigatable entity-action-network

For the ZENON project frames from the FrameNet[2] project [18] were used. These semantic frames are schematic representations of situation types (EATING, KILLING, SPYING, CLASSIFYING, etc.) together with lists of the kinds of participants, objects, and other conceptual roles that are seen as components of such situations. With the identified semantic frame the core and non-core frame elements are given.

We implemented semantic frames with transducers. If the system has recognized an action type it can determine the corresponding semantic frame. The semantic frame defines frame elements that have to be detected, i.e. the other parts of the sentence that are involved in the action (e.g., subject, object, time expressions, etc.). To fill in the frame elements, recognized named entities, POS tagging and expressions from the sentences are used. The filled-up semantic frames and other information from the processing of the natural language text are coded as typed feature structures in XML format. These structures represent the output of the natural language part of the ZENON system.

Finally, the knowledge extracted of different reports is combined and selected according to predefined XSLT (Extensible Stylesheet Language Transformation) sheets. These sheets implement the information needs of the analyst. The result is depicted graphically in an interactive network (see the graphs in Figure 2). The intelligence analysts can use this network to navigate through the meaning space of the reports.

[2] Online available at http://framenet.icsi.berkeley.edu/.

3 Applying Semantic Role Labeling for ZENONs Semantic Analysis

The ZENON project is concerned with automatic extraction of specific information from unrestricted text, i.e., extracting knowledge about actions and entities. The original semantic analysis of the system makes use of so called *action types* that have to be explicitly defined. This means, to be able to deduce an action type from certain text passages, the system needs rules that specify the textual context in which that action type can occur. Also, the semantic frame that is inferred from a recognized action type has to be manually encoded. Up to now this has been done only for a small selection of English verbs and semantic frames. To extent ZENONs coverage we realized a semantic role labeling (SRL) application that makes use of the lexical resource VerbNet (VN) [19]. The new SRL module is expected to improve the all-over performance of the ZENON system.

Semantic roles (also called *thematic roles*) are a form to represent the meaning of a text. They label the entities that are involved in an action with respect to the relations they have to each other in the context of that action. In the course of IE, semantic roles can be useful as they constitute further knowledge about actions and entities. When talking about semantic roles the linguistic concept of *verb argument structure* is important. A clause consists of a *main verb* and certain phrases (called *arguments*) that appear in a relationship with that verb. For every English clause the verb argument structure can be determined.

Verb arguments hold *syntactic relations* of subjects, objects and adverbial phrases. Every verb can occur with certain obligatory and facultative arguments. While a subject is always obligatory, it depends on the specific verb whether an object or an adverbial phrases are obligatory or facultative. For example *sleeping* has only one obligatory argument that is the subject (see *Henry* in sentences (3.1) and (3.2)). The verb like has two obligatory arguments as it needs at least one subject and one object (see sentence (3.3)). The adverbial phrase of location *in bed* in sentence (3.2) is an example for a facultative argument.

Sentence (3.1): [*Henry*]subject ***is sleeping***.

Sentence (3.2): [*Henry*]subject ***is sleeping*** [*in bed*]adverbial phrase of location.

Sentence (3.3): [*Henry*]subject ***likes*** [Lisa]object.

Semantic roles can be viewed as the *semantic level* of a verb's argument structure. Arguments are classified according to the semantic relation they have in the context of the action that is indicated by the verb. To give an example, the following sentences are annotated with semantic roles:

Sentence (3.4): [*Henry*]AGENT ***is sleeping***.

Sentence (3.5): [*Henry*]AGENT ***is sleeping*** [*in bed*]LOCATION.

Sentence (3.6): [*Henry*]AGENT ***likes*** [*Lisa*]THEME.

There is no consensus about a set of semantic roles or how exactly each role is defined. Some theories set semantic roles on a very abstract level, e.g. ARG0, ARG1, etc. (see (3.7)). Other approaches like FrameNet define very specific semantic roles for each verb, e.g. KILLER and VICTIM for the verb *kill* (see (3.8)). The lexical resource VerbNet defines 23 semantic roles on a medium level of generalization (see (3.9) for an example).

Sentence (3.7): [*The criminal*]ARG0 ***killed*** [*a man*]ARG1.

Sentence (3.8): [*The criminal*]KILLER ***killed*** [*a man*]VICTIM.

Sentence (3.9): [*The criminal*]AGENT ***killed*** [*a man*]PATIENT.

class	chase-51.6		
member verbs	chase, follow, pursue, shadow, tail, track, trail		
semantic roles	AGENT[+animate] THEME[+concrete] LOCATION		
frames	syntactic frame	semantic roles	example
	NP V NP	**NP**AGENT **V NP**THEME	*Jacky chased the thief.*
	NP V NP PP	**NP**AGENT **V NP**THEME **PREP NP**LOCATION	*Jacky chased the thief down the street.*
	NP V PP	**NP**AGENT **V after NP**THEME	*Jackie chased after the thief.*

Table 1: Schematic representation of VerbNet class chase-51.6

VerbNet[3] (VN) [19] is an online lexicon that provides syntactic and semantic information for more than 3700 English verbs. The verbs are hierarchically organized into *classes* and *subclasses* based on common syntactic and semantic features. See Table 1 for a schematic representation of a verb class. A verb can be member of more than one class due to semantic ambiguities. For example the verb *follow* is ambiguous and therefore member of four different verb classes: chase-51.6, comprehend-87.2-1, contiguous_location-47.8 and occurrence-48.3. Each VN verb class specifies one or more *verb frames* for its members. Verb frames define the syntactic context (*syntactic frame*), in which a certain verb can appear, i.e., the obligatory arguments a verb can take. Also the *semantic roles* are defined that the syntactic frame is associated with, i.e. what syntactic arguments of the verb can hold which semantic role. In this way VerbNet describes mappings from syntax to semantic, i.e., from syntactic frame to semantic roles, for each verb.

The syntax-semantic-mapping of VerbNet is elementary for our SRL approach. We apply this information to derive semantic roles from structural knowledge about the clause. The theoretic basis of our approach is the linguistic idea that there is a link between the syntactic structure of a text and its semantics. For a complete description of our approach see [5].

3.1 Architecture of the SRL application

The new SRL module can be run as a standalone application in GATE. It has not been integrated into the overall ZENON system, yet. The architecture of the SRL system (see Figure 3) consists of different processing components. The system accepts unrestricted text in English for input. Processing takes place sentence by sentence. As we base our approach on a mapping from syntax to semantics, first a syntactic analysis of the text is performed. The results are used to identify the main verb and its argument structure for each clause. Next, the system extracts matching VerbNet frames for the recognized verb argument structure. Finally, each clause is annotated with semantic roles that are indicated by its associated verb frame. Following the processing steps are described in more detail and illustrated by the use of an example (processing of sentence 3.10).

Sentence (3.10): *The suspect is following the politician into a public building*.

[3] Online available at http://verbs.colorado.edu/~mpalmer/projects/verbnet.html.

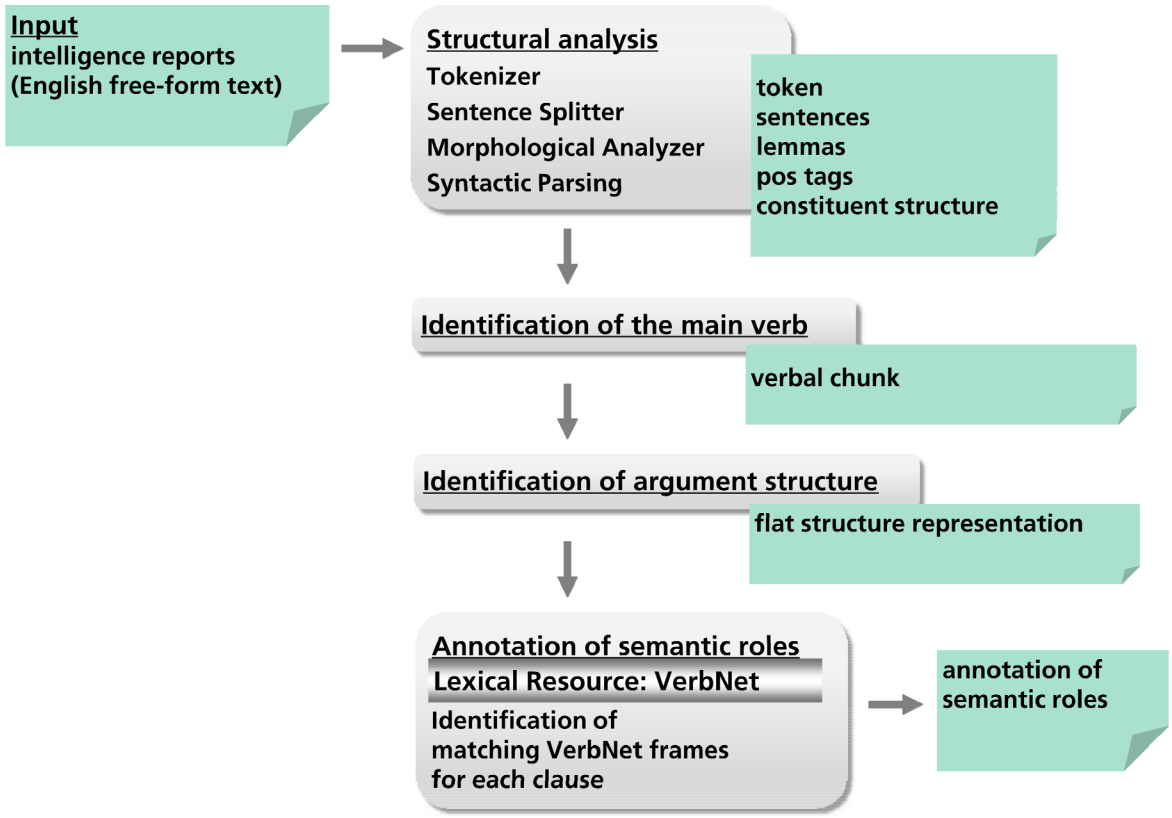


Figure 3: Processing chain of the SRL system

Structural Analysis

Structurally analyzing the text is important as preprocessing, to identify syntactic structure of sentences, and to receive a lemma form for each verb. Figure 4 shows some of the annotations that result from structural analysis. The text is tokenized and sentence splitting is done (see *token* and *sentence* annotations in Figure 4). A tokenizer and a sentence splitter are provided by GATE software. GATE's morphological analyzer determines the lemma (dictionary form) for every word (see *lemma* annotations in Figure 4). For example the word *following* is annotated with its lemma *follow*.

Every sentence is pos tagged and parsed (see *syntax-tree-nodes* and *dependencies* in Figure 4). We use the *Stanford Parser*[4] [20; 21], a statistical parser that outputs dependency structures [22] (see Figure 5) as well as constituent trees (see Figure 6). GATE comes with a plugin that acts as a wrapper around the Stanford Parser, which makes integration easy. Constituent trees and dependency structures are two different ways of representing a sentence's syntactic structure. They will later in the processing chain be of use for the identification of verb argument structure.

[4] Online available at http://nlp.stanford.edu/software/lex-parser.shtml

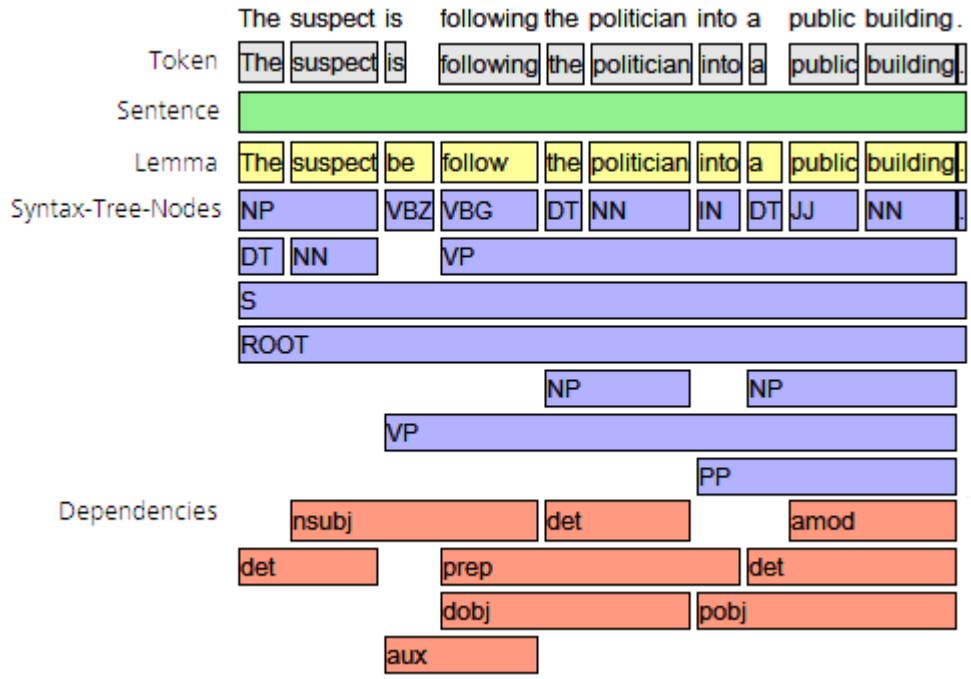


Figure 4: Example of GATE annotation resulting from structural analysis

syntax-tree-node and dependency annotations are not sorted by their associated tree level

dependency annotations stretch in between the related tokens

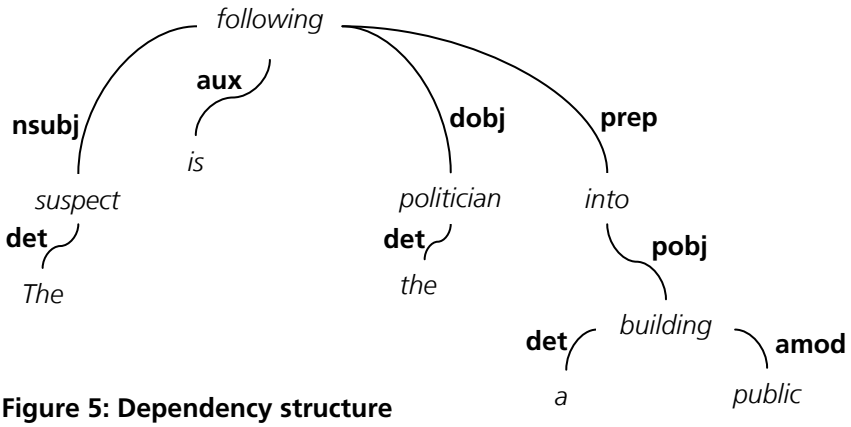


Figure 5: Dependency structure

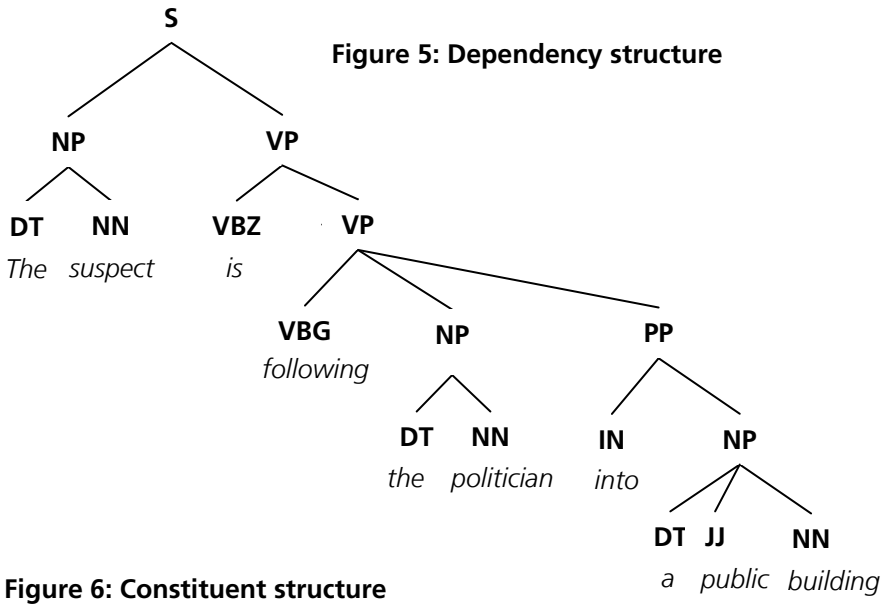


Figure 6: Constituent structure

root- and .-node omitted

Identification of verb argument structure

Extracting the verb argument structure of a clause is an essential step in semantic role labeling. Arguments are labeled with semantic roles due to a semantic frame that is determined by the main verb. We developed a component that identifies the parts of the sentence (i.e. sequences of tokens) that constitute the main verb and its arguments. It is implemented in Java. The output is a flat structural representation for each clause depicting the extracted verb argument structure.

The system uses the results of syntactic parsing to identify the verb argument structure for each clause. First, dependency annotations are taken into account. Stanford dependencies describe binary syntactic relations between the words of a clause starting with the main verb (see Figure 5). For example in sentence (3.10) the verb *following* has a subject, a direct object, a prepositional phrase and an auxiliary which are represented by the following dependency annotations (see Figure 4 and Figure 5):

nsubj (*following* , *suspect*)

dobj (*following* , *politician*)

prep (*following* , *into*)

aux (*following* , *is*).

Based on those relations, the main verb and certain words being connected to that verb are grouped together as the verbal chunk. The main verb is derived from the **nsubj** dependency, because a subject must always be an obligatory argument of a main verb. Via further heuristics that use the dependency annotations the system infers which other words belong to the main verb. E.g., Figure 7 shows how the auxiliary *is* and the main verb *following* are annotated as the verbal chunk *is following*.

The system applies the main verb's dependency annotations to identify its arguments. Thus *following* takes three arguments, including a subject, an object and a prepositional phrase. Arguments can comprise single words or sequences of words (i.e., *phrases* or *chunks*). Stanford dependencies describe relations between single words only. We therefore need to determine a chunk for every such argument's relation. This is done on the basis of the constituent structures that are output by the Stanford Parser. For every identified verb argument a syntax tree node is extracted that comprises a fitting chunk. E.g., for the argument *suspect*, which has a **nsubj** relation with *following*, the phrase *the suspect* with the node **NP** is selected (see Figure 6). In this way a flat syntactic representation of each clause is generated (see *phrase structure* annotation in Figure 7).

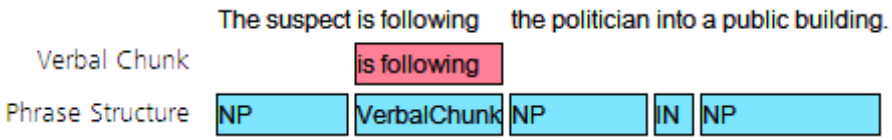


Figure 7: Annotation of verbal chunks and phrases

Tags: NP = nominal phrase, VerbalChunk = verbal chunk, IN = preposition

Annotation of semantic roles

The last processing component of the SRL system annotates the identified verb arguments of each clause with semantic roles. The semantic role information is extracted from the associated VerbNet frame. We find VN frames on the basis of the identified verb argument structure.

For each clause we extract all VerbNet frames that are associated with the main verb. We know the main verb of each clause (see *verbal chunk* annotation in Figure 7) and its dictionary form (see lemma annotation in Figure 4) from previous processing steps. For example in sentence (3.10) following is identified as main verb and its lemma form is follow. For the verb follow VerbNet defines altogether 19 frames in 4 verb classes. Each verb frame consists of a specific

syntactic structure and its associated semantic roles. The format of the syntactic structure resembles the flat syntactic representation we generated for each clause. In that way we are able to determine all frames with a matching syntactic structure. For the verb follow there is only one VerbNet frame with a syntactic structure that matches the extracted syntactic representation of sentence (3.10). This frame is defined by the class *chase-51.6* (see Table 1). It describes the following syntactic structure:

NP Verb NP Preposition NP

From that matching VerbNet frame we extract the semantic role information. In case of the example frame this is:

AGENT VERB THEME PREPOSITION LOCATION.

Finally, each clause is annotated with semantic roles, i.e. the verb argument structure that has previously been determined is annotated with the semantic roles that are defined by the matching VerbNet frame (see Figure 8).

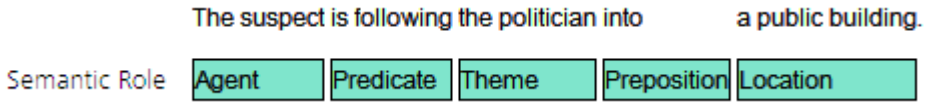


Figure 8: Annotation of semantic roles

3.2 Open Problems

VerbNet's verb coverage can cause problems when the identified verb is not existent in the resource. In this case the system will not receive any verb frame and no semantic roles can be extracted. To extend VerbNet's coverage other lexical resources like FrameNet could be taken into account or hybrid systems that take statistical approaches into account might be tested. A further challenge is disambiguation of verb frames, when there is more than one matching verb frame for a clause. Here too, different lexical resources, or statistical approaches could be helpful.

The structural analysis is essential for the SRL process. Due to problems of statistical parsing approaches the identified syntactic structure may be erroneous. As a result, semantic analysis may be wrong or fail. Different syntactic parsing techniques need to be tested to see whether structural analysis can become more robust.

4 Conclusion

Processing of human language is identified as a critical capability in many future military applications. Our research system ZENON is performing a partial content analysis of English free-form texts. In this paper, we described how to extend the semantic analysis of ZENON by applying a semantic role labeling approach. To suit the military domain we realized a non-statistical SRL-approach. For each action encoded in the text the verb and the participating entities are extracted and their semantic roles are identified. We use a statistical syntactic parser to generate a formalized syntactic representation of each sentence. We then find the matching semantic roles from the lexical resource VerbNet. At the moment, we are in the process of integrating the semantic module into the current ZENON system. We expect that systems like ZENON will increase productivity of the intelligence analyst.

References

[1] **Pigeon, S., et al.** *Use of Speech and Language Technology in Military Environments.* 2003. NATO Technical Report, TRIST037.

[2] **Steeneken, H. J. M.** *Potentials of Speech and Language Technology Systems for Military Use: an Application and Technology Oriented Survey.* 1996. NATO, Technical Report, AC/243(Panel 3)TP/21.

[3] **Hecking, M. and Sarmina-Baneviciene, T.** A Tajik Extension of the Multilingual Information Extraction System ZENON. *Proceedings of the 15th International Command and Control Research and Technolgy Symposium (ICCRTS), Santa Monica, CA, USA, June 2010.*

[4] **Hecking, M.** Multilinguale Textinhaltserschließung auf militärischen Texten. [Hrsg.] Michael Wunder und Jürgen Grosche. *Verteilte Führungsinformationssysteme.* Heidelberg, Germany : Springer, 2009.

[5] **Noubours, S.** *Annotation semantischer Rollen in HUMINT-Meldungen basierend auf dem statistischen Stanford Parser und der lexikalischen Ressource VerbNet. FKIE-Bericht Nr. 195.* Wachtberg, Germany : Fraunhofer-FKIE, 2010.

[6] **Jurafsky, D. und Martin, J. H., [Hrsg.].** *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.* 2nd Edition. Upper Saddle River, NJ, USA : Prentice Hall, 2009.

[7] **Jurafsky, D.** Semantic Role Labeling, Lecture CS 224U, Winter 2010. [Online] http://www.stanford.edu/class/cs224u/lec/224u.10.lec5.pdf.

[8] **Schwerdt, C.** *Analyse ausgewählter Verbalgruppen der Sprache Dari zur multilingualen Erweiterung des ZENON-Systems. FKIE-Bericht Nr. 146.* Wachtberg, Germany : Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), 2007.

[9] **Hecking, M. and Schwerdt, C.** Multilingual Information Extraction for Intelligence Purposes. *Proceedings of the 13th International Command and Control Research and Technolgy Symposium (ICCRTS), Bellevue, WA, USA, June 2008.*

[10] **Sarmina-Baneviciene, T.** *Analyse spezifischer Probleme der tadschikischen Sprache zur multilingualen Erweiterung des ZENON-Systems. FKIE-Bericht Nr. 196.* Wachtberg, Germany : Fraunhofer FKIE, 2010.

[11] **Hecking, M.** *Das KFOR-Korpus als Ergebnis semantisch annotierter militärischer Meldungen, FKIE-Bericht Nr. 124.* Wachtberg, Germany : Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), 2006.

[12] **Hecking, M.** The KFOR Text Corpus. *Proceedings of the 12th International Command and Control Research and Technolgy Symposium, Newport, USA, June 2007.*

[13] **Hecking, M.** Analysis of Free-form Battlefield Reports with Shallow Parsing Techniques. *RTO IST Symposium on „Military Data and Information Fusion", Prague, Czech Republic, October 2003.*

[14] **Hecking, M.** How to Represent the Content of Free-form Battlefield Reports. *Proceedings of the 2004 Command and Control Research and Technology Symposium (CCRTS) "The Power of Information Age Concepts and Technologies", San Diego, California, USA, June 2004.*

[15] **Appelt, D. and Israel, D.** Introduction to Information Extraction Technology. Stockholm: IJCAI-99 Tutorial. [Online] 1999. http://www.ai.sri.com/~appelt/ie-tutorial/.

[16] **Cunningham, H., et al.** GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.*

[17] **Cunningham, H., et al.** Developing Language Processing Components with GATE. [Online] 2010. http://gate.ac.uk/sale/tao/tao.pdf.

[18] **Ruppenhofer, J., et al.** FrameNet II: Extended theory and practise. [Online] 2010. http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126.

[19] **Kipper, K., et al.** A largescale classification of english verbs. *Language Resources and Evaluation.* 42 (1), March 2008, S. 21-40.

[20] **Klein, D. and Manning, C. D.** Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics, Sapporo, Japan, July 2003.*

[21] **Klein, D. and Manning, C. D.** Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in Neural Information Processing Systems 15, Proceedings of the 16th Annual Conference on Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada, December 2002.*

[22] **de Marneffe, M.-C., MacCartney, B. and Manning, C. D.** Generating typed dependency parses from phrase structure parses. *In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), Genua, Italy, Mai 2006.*

Semantic Analysis of Military Relevant Texts for Intelligence Purposes

Sandra Noubours
Dr. Matthias Hecking
Fraunhofer Institute for Communication,
Information Processing and Ergonomics FKIE
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven

sandra.noubours@fkie.fraunhofer.de
matthias.hecking@fkie.fraunhofer.de

Fraunhofer FKIE

Outline

- Introduction
- The ZENON System
  and its information extraction functionalities
- A Semantic Role Labeling application
  for ZENONs Semantic Analysis
- Conclusion

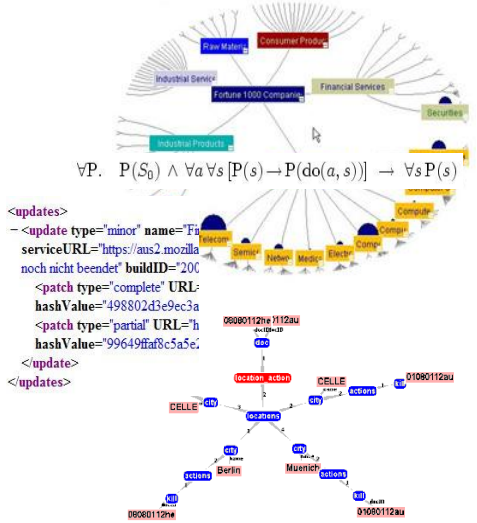
1. Introduction

Military intelligence

For military intelligence, large quantities of intelligence reports and other documents need to be analyzed.

content analysis

What?
Who? Where?
When? …?



Text from different sources can contain military relevant information (HUMINT, OSINT, …).

Information need to be extracted.

1. Introduction

NLP for military intelligence

Natural language processing (NLP) can be applied to efficiently handle analysis of textual data.

➡ We set up the research project **ZENON**.

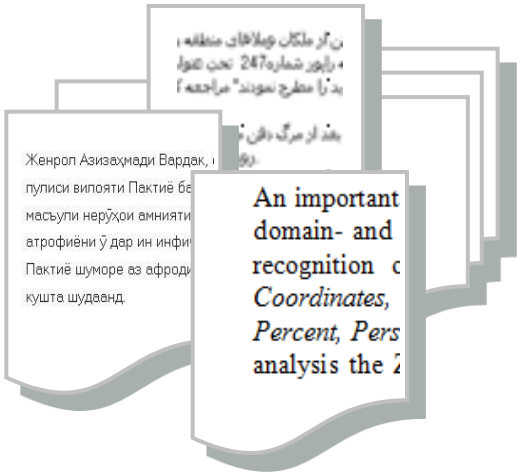
- ZENON realizes a (prototypical) **information extraction (IE) system** for the (partial) content analysis of English HUMINT reports.
- The system has further been extended for **multilingual information extraction**, i.e., processing Dari and Tajik texts.
- Here, we present the **improvement** of ZENON's English **semantic analysis** by **semantic role labeling (SRL)**.

THE ZENON SYSTEM

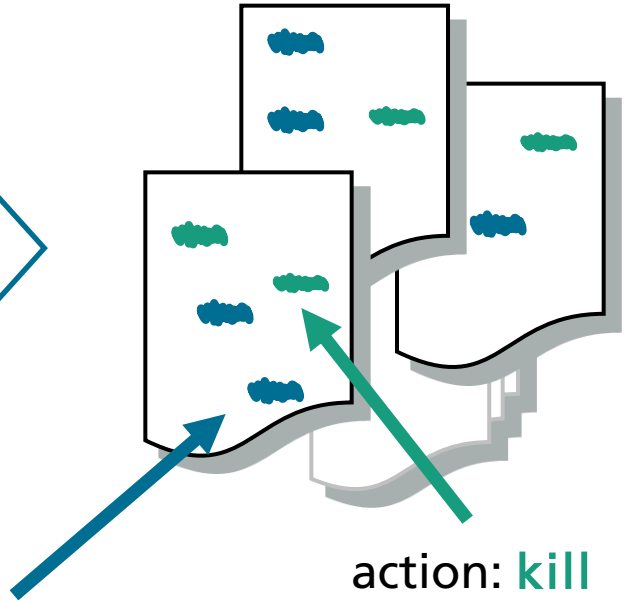
2. The ZENON System

Information extraction

Information extraction (IE) means the (partial) **content analysis** of free-form text. **Relevant information** about a specific entity and/or action in natural language texts is **identified, extracted and represented ...**



identification of relevant information



action: **kill**

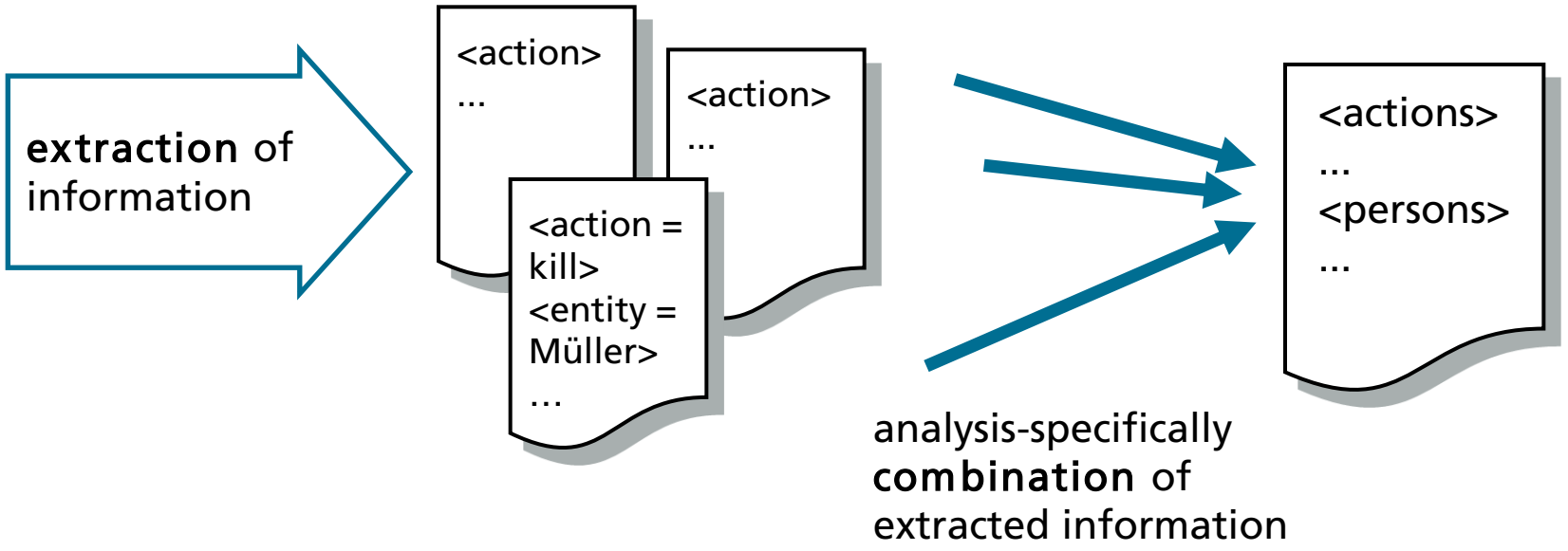
entity: **MUELLER**

Knowledge about **actions** and **entities** is distributed over many reports.

2. The ZENON System

Information extraction

Information extraction (IE) means the (partial) **content analysis** of free-form text. **Relevant information** about a specific entity and/or action in natural language texts is **identified, extracted and represented ...**

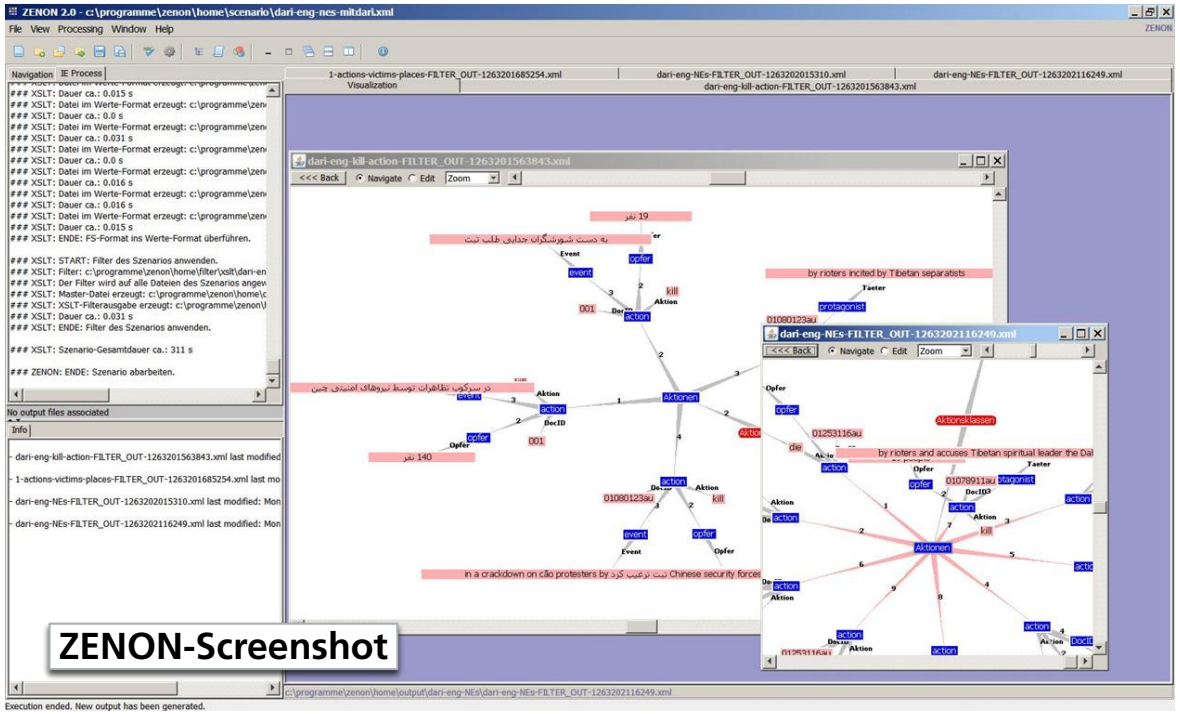


2. The ZENON System

Information extraction

Information extraction (IE) means the (partial) **content analysis** of free-form text. **Relevant information** about a specific entity and/or action in natural language texts is **identified, extracted and represented ...**

graphical **representation**



2. The ZENON System

Architecture

implemented using GATE software

Input: intelligence reports

Structural analysis

- Token Annotation
- POS Tags
- Sentence Annotation
- Named Entities
- Morphological Information & Lemma

Detection of verb phrases

- Verbal Chunks

Semantic Analysis

Recognize action types
→ **Infer semantic frames**

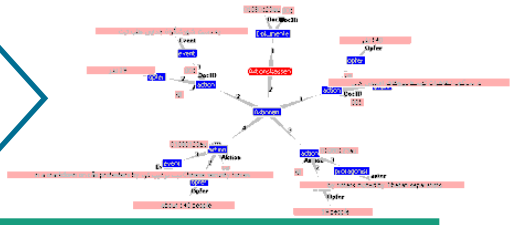
Extract relevant information
→ **Fill semantic frames**

- Semantic Frames (feature structures)

Select extracted information

Combine extracted information

Visualization



2. The ZENON System

Original semantic processing

ZENON's **original semantic analysis** is based on **semantic frames** that are inferred from **action types**.

Example: *John Mueller was killed in an explosion incident.*

a) semantic analysis follows structural analysis and identification of verb phrases
[*John Mueller*] [*was killed*] [*in*] [*an explosion incident*].

b) the system is able to deduce an **action type** from certain text passages
… *was killed* … ➡ action type *KILL*: **A** *was killed in* **B**

c) the recognized action type determines the **semantic frame**
A *was killed in* **B** ➡ semantic frame: **VICTIM** *was killed in* **CAUSE**

d) the semantic frame is **filled with the identified entities**
[*John Mueller*]VICTIM was killed *in* [*an explosion incident*]CAUSE.

2. The ZENON System

Original semantic processing: Problems

Action types & **semantic frames** have to be **explicitly defined**.

- manually written grammar rules specify in what textual context an action type can occur
- a semantic frame (that is inferred from a recognized action type) has to be manually encoded

Such a **manual approach is time-consuming and inefficient**.

➡ has been realized only for a small selection of English verbs and semantic frames

➡ **low coverage**

To **improve ZENON's semantic analysis**, we realized a **semantic role labeling** (**SRL**) application.

Improving ZENONs semantic analysis by an

SRL APPLICATION

3. SRL for ZENON

Semantic roles

Semantic roles are a form to represent the meaning of a text.
In the **context of a specific action**, **each entity** involved has a **certain semantic role**.

[*The policemen*]AGENT arrested [*the suspect*]PATIENT.

[*John Mueller*]VICTIM was killed *in* [*an explosion incident*]CAUSE.

3. SRL for ZENON

Semantic Role Labling

Semantic role labeling (**SRL**) is the process of automatically indentifying semantic roles in a text. For each action encoded in a text the participating entities are identified and labeled with semantic roles.

In the course of **Information Extraction**, SRL can be useful as semantic roles constitute further knowledge about actions and entities.

3. SRL for ZENON

Approach

Different SRL approaches exist, many of them are applying **machine learning** (statistical approach) where the system is trained on an annotated corpus.

- the training corpus should be domain specific
- there is no such corpus annotated with semantic roles existent for the military domain

➡ we implemented a **non-statistical approach** that makes use of a lexical resource

3. SRL for ZENON

Approach

Our SRL approach is based on a **syntax-semantic-mapping**

- we derive **semantic roles from structural knowledge** about the clause
- for this purpose we apply information from the **lexical resource VerbNet** (VN)
  - VerbNet is an online lexicon that provides syntactic and semantic information for more than 3700 English verbs
  - semantic roles are defined in the context of syntactic structures
  - in this way **VerbNet describes mappings from syntax to semantic**, i.e., from syntactic frame to semantic roles, for each verb

3. SRL for ZENON

Processing

For each action encoded in the text the verb and the participating entities are extracted (*syntactic analysis: a&b*) and their semantic roles are identified (*semantic annotation: c&d*). [➡ *syntax-semantic-mapping*]

a) **structural analysis** of the text is performed:
applying a statistical syntactic parser to generate a formalized syntactic representation of each sentence

b) results are used to identify the **main verb and its argument structure** for each clause

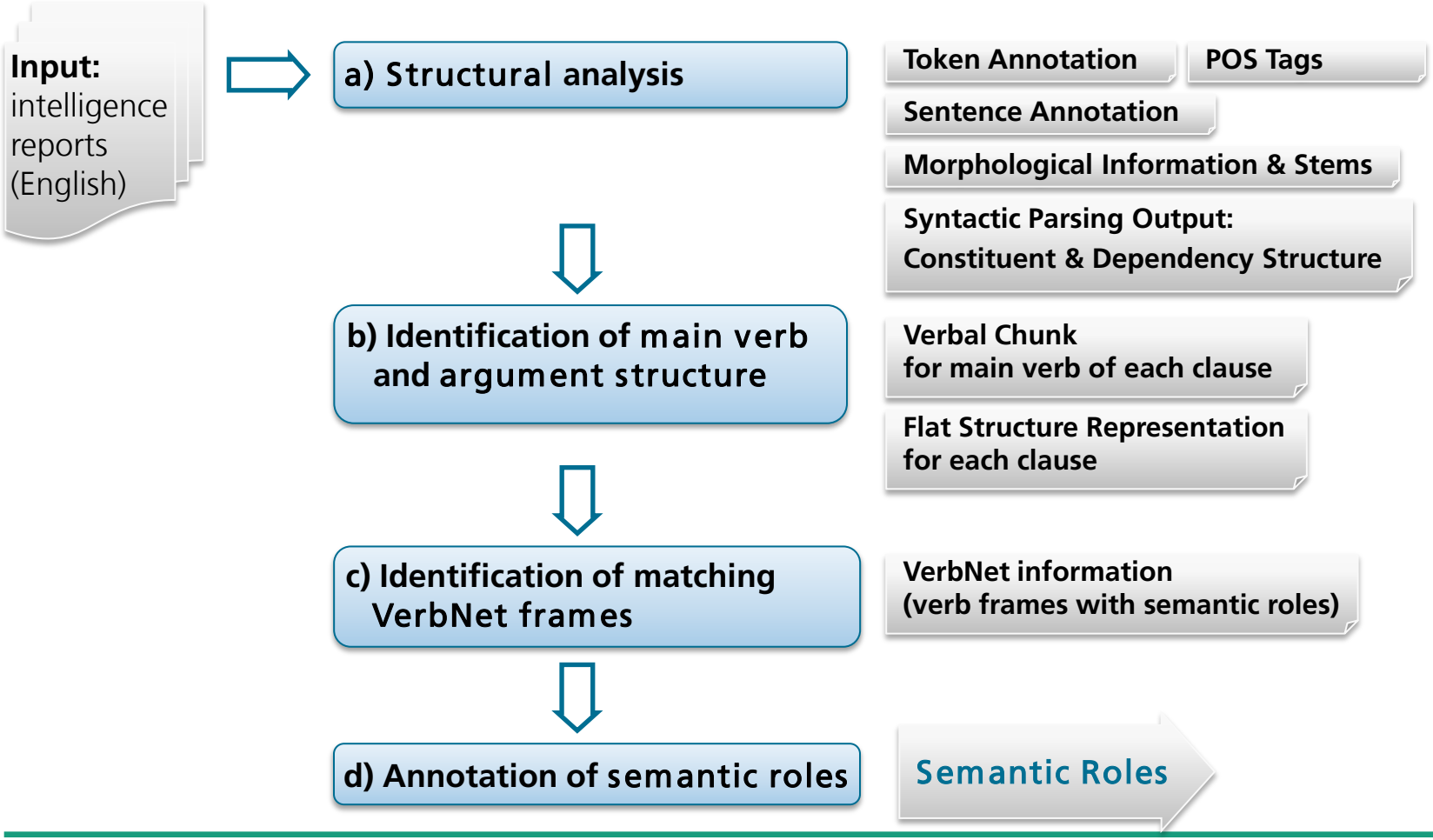
c) for each recognized main verb and its verb argument structure, the system **extracts matching VerbNet information** which also include semantic roles

d) finally, the clause (i.e., the identified verb argument structure) is annotated with the **semantic roles**

3. SRL for ZENON

Architecture

implemented using GATE software



3. SRL for ZENON

Processing step a) structural analysis

Tokenization, **sentence splitting** and **morphological analysis**…

Example: *The suspect is following the politician into a public building .*

Token – Annotation:

The	suspect	is	following	the	politician	into	a	public	building	.

Sentence – Annotation:

The suspect is following the politician into a public building .

Lemma – Annotation:

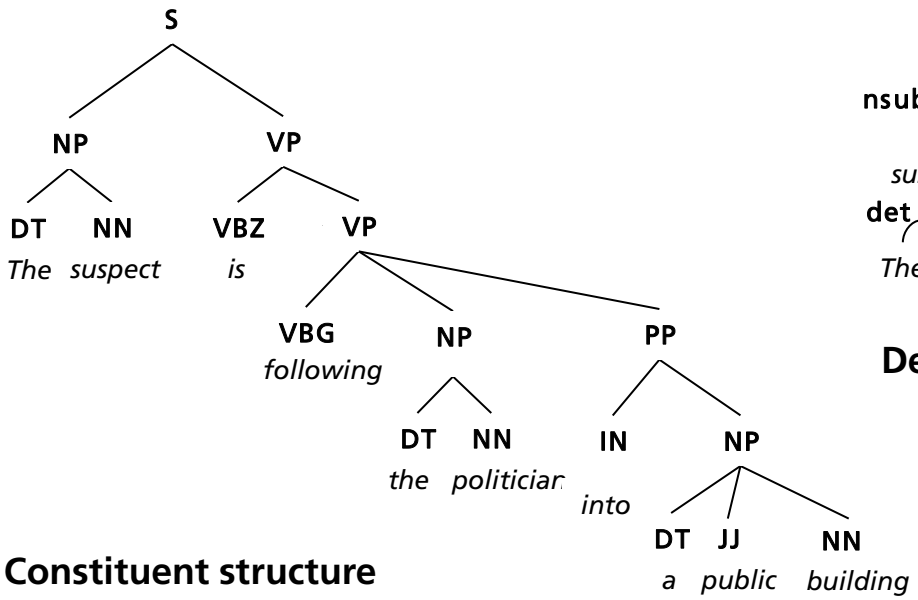
The	suspect	be	follow	the	politician	into	a	public	building	.

3. SRL for ZENON

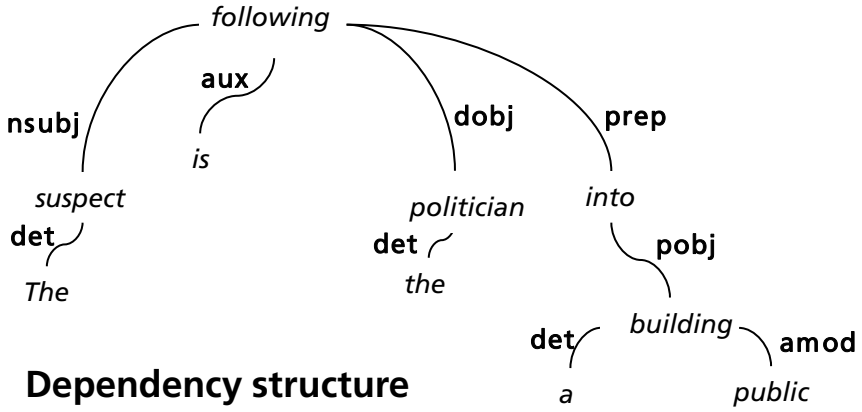
Processing step a) structural analysis

… **syntactic parsing** (Stanford Parser) of
constituent and dependency structure for each sentence.

Example: *The suspect is following the politician into a public building .*



Constituent structure



Dependency structure

3. SRL for ZENON

Processing step b) identification of argument structure

Based on information from structural analysis:
Identification of **main verb** and **argument structure** for each clause.

Example: *The suspect is following the politician into a public building .*

Verbal Chunk – Annotation:

is following

chunk of main verbal expression
(main verb + auxiliaries + modal verbs + …)

Phrase Structure – Annotation:

NP	Verbal Chunk	NP	IN	NP

flat syntactic representation of argument structure
(Tags: NP = nominal phrase, Verbal Chunk = chunk of main verb, IN = preposition)

3. SRL for ZENON

Processing step c) extraction of VerbNet information

Extraction of matching **VerbNet information**.

Example: *The suspect is following the politician into a public building .*

- get **lemma** for main verb

Verbal Chunk – Annotation: is following

Lemma – Annotation: follow

- convert representation of recognized argument structure of each clause into format of **VerbNet syntactic frames**

Phrase Structure – Annotation:

NP	Verbal Chunk	NP	IN	NP

Phrase Structure (in VerbNet format):

NP	Verb	NP	Preposition	NP

3. SRL for ZENON

Processing step c) extraction of VerbNet information

Extraction of matching **VerbNet information**.

Example: *The suspect is following the politician into a public building .*

Lemma of main verb: **follow**

+

Phrase Structure (VerbNet format):

NP **Verb** **NP** **Preposition** **NP**

VerbNet

matching VerbNet verb frame (VerbNet class chase-51.6)**:**

[NP]AGENT **Verb**: **follow** **[NP]THEME** **Preposition** **[NP]LOCATION**

VerbNet – Semantic Roles:

AGENT **THEME** **LOCATION**

3. SRL for ZENON

Processing step d) annotation with semantic roles

Annotation of verb-argument-structure with **VerbNet semantic roles**.

Example: *The suspect is following the politician into a public building .*

Verb-Argument-Structure (Phrase Structure – Annotation):

NP	Verbal Chunk	NP	IN	NP

\+

VerbNet – Information:

[NP]AGENT Verb: follow [NP]THEME Preposition [NP]LOCATION

AGENT	THEME	LOCATION

Semantic Role – Annotation:

AGENT	THEME	LOCATION

[*The suspect*]AGENT *is following* [*the politician*]THEME *into* [*a public building*]LOCATION .

ZENON

CONCLUSION

4. Conclusion

- **Processing of human language** is a critical capability in many future **military applications**.
- **ZENON** is a prototypical **information extraction system** for the partial content analysis of free-form texts. We expect that systems like ZENON will **increase productivity of the intelligence analyst**.
- We implemented a SRL application to improve ZENON's semantic analysis. This is expected to **improve the all-over performance of the ZENON system**.

Reference

- S. Noubours. *Annotation semantischer Rollen in HUMINT-Meldungen basierend auf dem statistischen Stanford Parser und der lexikalischen Ressource VerbNet*. FKIE-Bericht Nr. 195. Wachtberg, Germany: Fraunhofer-FKIE, 2010.
- M. Hecking and T. Sarmina-Baneviciene. *A Tajik Extension of the Multilingual Information Extraction System ZENON*. In: Proceedings of the 15th International Command and Control Research and Technolgy Symposium (ICCRTS), Santa Monica, CA, USA, June 2010.
- M. Hecking. *System ZENON – Semantic Analysis of Intelligence Reports*. In: Proceedings of the LangTech 2008, February 28-29, 2008, Rome, Italy
- M. Hecking. *Multilinguale Textinhaltserschließung auf militärischen Texten*. In: M. Wunder and J. Grosche (Ed.). Verteilte Führungsinformationssysteme. Heidelberg, Germany: Springer, 2009.

Thank you for your attention!

Questions?